

An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

Ryan Giordano (rgiordan@mit.edu)¹
January 2022

¹With coauthors Rachael Meager (LSE) and Tamara Broderick (MIT)

Dropping data: Motivation

More data & cheaper computation \Rightarrow

Statistical analyses are playing larger roles in decision making.

Decisions are important: We want **trustworthy** conclusions.

Data / models not always perfect: We want **robust** conclusions.

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Running example: Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit based on 16,560 data points.

We can reverse the studies qualitative conclusions by removing 15 observations ($< 0.1\%$ of the data).

How do we find sets of influential points? Difficult in general!

We provide a **automatic approximation** with finite-sample guarantees.

Studying the approximation reveals the causes of non-robustness.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Beta (SE)
Original result	-4.55 (5.88)

Original conclusion:

There is no evidence that microcredit is effective.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)

Original conclusion:

There is no evidence that microcredit is effective.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)

Original conclusion:

There is no evidence that microcredit is effective.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change sign and significance	15	7.03 (2.55)

Original conclusion:

There is no evidence that microcredit is effective.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change sign and significance	15	7.03 (2.55)

Original conclusion:

There is no evidence that microcredit is effective.

Potential conclusions after data dropping:

The effect of microcredit is positive (negative) & statistically significant.

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change sign and significance	15	7.03 (2.55)

Original conclusion:

There is no evidence that microcredit is effective.

Potential conclusions after data dropping:

The effect of microcredit is positive (negative) & statistically significant.

The culprit is signal to noise ratio.

By the end of the talk, we will see that the sensitivity is due to

- High variability of the outcome (household profit) relative to
- A small signal driving the conclusion (statistical significance)

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Not always! But sometimes, surely yes.

Thinking without random noise can be helpful.

Suppose you have a farm, and want to know whether your average yield is greater than 170 bushels per acre. At harvest, you measure 200 bushels per acre.

- Scenario one: If your yield is greater than 170 bushels per acre, you make a profit.
 - Don't care about sensitivity to small subsets
- Scenario two: You want to recommend your farming methods to a friend across the valley.
 - Might care about sensitivity to small subsets

For example, often in economics:

- Small fractions of data are missing not-at-random,
- Policy population is different from analyzed population,
- We report a convenient summary (e.g. mean) of a complex effect,
- Models are stylized proxies of reality.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

The number of subsets $\binom{N}{\lfloor \alpha N \rfloor}$ can be very large even when α is very small.

In the MX microcredit study, $\binom{16560}{15} \approx 1.4 \cdot 10^{51}$ sets to check for $\alpha = 0.0009$.

We provide a fast, automatic approximation based on the **influence function**.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Non-robustness to removal of $\lfloor \alpha N \rfloor$ points is:

- Not (necessarily) caused by misspecification.
- Not (necessarily) caused by outliers.
- Not captured by standard errors.
- Not mitigated by large N .
- Primarily determined by the **signal to noise** ratio
... in a sense which we will define.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

- We provide deterministic error bounds for small α .
- We show the accuracy in simple experiments.
- We show the accuracy in a number of real-world experiments.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Conclusion: Related work and future directions

Question 1:

How do we find influential datapoints?

Question 2:

What makes an estimator non-robust?

Question 3:

When is our approximation accurate?

Conclusion:

Related work and future directions

Tamara Broderick, Ryan Giordano, Rachael Meager (alphabetical authors)
“An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?”

<https://arxiv.org/abs/2011.14999>

M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.